基于 LDA 挖掘计算机科学文献的研究主题

杨海霞　高宝俊　孙含林
(武汉大学经济与管理学院　武汉 430072)

摘要: 【**目的**】运用文本挖掘技术自动从海量科技文献中提取研究主题并探测其研究趋势。【**方法**】以《中文核心期刊要目总览(2014 年版)》—"TP 自动化技术、计算机技术"栏目前 10 种期刊刊载的计算机科学类(Computer Science)文献为研究对象, 借助 LDA 主题模型, 考虑科技文献的发表时间信息, 挖掘出典型话题, 并根据主题强度分析主题的演化趋势。【**结果**】18 个研究话题中有 7 个主题强度上升的主题和 6 个主题强度下降的主题。【**局限**】仅分析了国内计算机领域的前 10 种期刊, 期刊范围不够大, 也未考虑国外计算机领域的期刊文献。【**结论**】该方法能够深入挖掘计算机领域期刊文献的话题, 帮助从事该领域研究的学者了解主题的演化趋势并寻找新兴研究主题。

关键词: 计算机科学　LDA　主题提取　主题强度　文档聚类
分类号: G350

1　引　言

计算机科学是系统性研究信息与计算的理论基础以及它们在计算机系统中如何实现与应用的实用技术的学科。在信息技术高速发展的时代, 计算机科学已成为各个国家不可或缺的学科领域。而科技文献作为学术成果的重要载体, 凝聚了科研人员的大量智慧, 是传播知识、进行学术交流的重要途径。因此, 探测计算机科学类文献的研究内容, 能够了解计算机科学领域的发展状态。

笔者借助概率主题模型 LDA(Latent Dirichlet Allocation)[1-3], 对入选《中文核心期刊要目总览(2014 年版)》—"TP 自动化技术、计算机技术"栏目前 10 种期刊刊载的计算机科学(Computer Science)文献进行文本建模, 结合困惑度和专家判断确定模型的最优主题数, 同时考虑文献的发表时间信息, 从主题内容和主题强度两方面, 探测 2006 年–2015 年期间国内计算机科学领域中各个研究主题的发展趋势, 并根据 LDA 模型结果对文档进行聚类, 统计各个主题下的文献数量, 深入了解热点主题下的文档内容。通过本文分析, 科研人员能够初步了解国内计算机科学领域近 10 年的研究状态, 把握计算机科学领域的新兴研究主题, 并判断哪些主题值得继续研究。

2　相关研究现状

近年来, 为把握计算机科学的发展状态, 相关学者从计算机科学领域的整体或部分的角度, 对计算机科学的发展过程进行论述。如, 郭玉等[4]采用文献计量学和科学计量学的研究方法, 从论文的时间分布、被引用情况及主题分布等方面, 分析中国作者在国际期刊发表的计算机科学论文, 以期了解计算机科学的学科现状; 陈国良等[5]综述了并行计算的一体化研究现状, 并展望了其发展趋势; 章锦文等[6]以叙述的方式, 讨论了神经网络计算机的研究现状和发展趋势。

上述研究中, 或以文献数量统计的方法, 或以文献综述的方法, 少有学者借助主题模型分析计算机科学类文献的研究主题, 以适应当前文献数量巨增的现状。本文借助 LDA 模型, 抽取 29 621 篇计算机科学文献的研究主题, 并根据不同时段的主题强度, 深入分析计算机科学的主要研究主题及各个主题的发展趋势。

LDA 模型是生成式概率主题模型, 假定潜在主

通讯作者: 高宝俊, ORCID: 0000-0001-0001-5146-3740, E-mail: 18064034195@163.com。

题是语料中一系列词的概率分布，文档是一系列潜在主题的概率分布。在 LDA 模型中，通常同一个主题中的词存在高语义相关性，如在主题“安全密钥”中，词“secure”“scheme”“key”“protocol”“signature”“authentic”均与该主题高度相关。此外，与一般的聚类方法不同，LDA 允许一个文档同时包含多个主题，故更适用于提取科技文献的研究主题[1-3]。

国外研究中，Griffiths 等[3]首先将 LDA 模型用于提取 *PNAS* 期刊文献摘要的主题及主题变化趋势，并用 Gibbs 抽样算法推断 LDA 模型。随后，LDA 模型被陆续用于分析生物医学[7]、计算机语言学[8]、文献计量学[9]、图书信息管理学[10]、经济学[11]等领域的科技期刊文献，自动挖掘大量文献的研究主题，了解某个领域的研究状态。

国内相关学者也多应用 LDA 模型进行科技文献的情报分析，如贺亮等[12]对 NIPS 论文集和 ACL 论文集进行实验，用主题词的类 TF-IDF 值，探讨主题的内容演化过程；关鹏等[13]对不同语料库下的 LDA 主题模型进行对比研究，并对主题抽取效果进行评价；李湘东等[14]在 LDA 模型中引入时间因素，以探测科技期刊的主题演化；王曰芬等[15]以知识流领域为研究对象，借助 LDA 挖掘不同学科下的知识流研究结构。此外，为满足不同的科技情报分析需求，许多学者对 LDA 模型进行改进，如王萍[16]串联文献的文本信息和作者信息，构建主题–作者(Topic-Author)的模型；叶春蕾等[17]综合科研文献的关键词和引文，构建引文–主题概率模型；王平[18]考虑文献发表的时间和题录信息，构建分层 LDA 模型，找到热点话题以及话题的演化特性；王金龙等[19]针对目前科研文献主题演化概率分布问题，阐述了主题与事件的关联关系，提出一种新型的基于模块化的主题方法；李湘东等[20]将 SVM 算法加入 LDA 模型中，优化主题分类；秦晓慧等[21]在 LDA 模型中加入主题关联过滤规则，以期减少非关联主题的干扰问题；杨如意等[22]基于 LDA 模型，融合作者和时间两个外部特征，以展示文档内容、主题和作者之间的动态关系。

综上，笔者借助 LDA 经典模型，对计算机科学文献进行主题抽取，并对各个主题的内容和强度进行细致分析，以期深入了解我国计算机科学在 2006 年–2015 年期间的研究状态。

3　数据与实验

3.1　数据来源

以《中文核心期刊要目总览(2014 年版)》为基准，选取“TP 自动化技术、计算机技术”学科中排名前 10 的期刊文献为目标样本，对计算机科学领域的文献话题进行提取和分析。研究数据来自中国科学引文数据库，具体检索策略为：出版物名称=“计算机学报” OR “软件学报” OR “自动化学报” OR “计算机研究与发展” OR “控制与决策” OR “中国图象图形学报” OR “系统仿真学报” OR“计算机辅助设计与图形学学报” OR “计算机应用” OR “计算机科学”，时间跨度=2006-2015，研究方向=“Computer Science”，文献类型=“Article”，选取字段“英文标题(TI)、英文关键字(DE)、英文摘要(AB)、来源期刊(SO)和发表时间(PY)”，得到 31 983 条记录。由于本文试图通过分析科技文献的摘要来提取主题，因此首先需要删除前言、致谢等非科技文献，其次删除标题、关键词、摘要不完整的文献，最后获得 29 621 条文献记录。

2006 年–2015 年期间，样本中各个期刊的计算机科学类文献占比如图 1 所示。占比较大的依次是《计算机应用》(24%)、《计算机科学》(22%)、《系统仿真学报》(16%)，这是由于这三个期刊的发文数量一直较大。

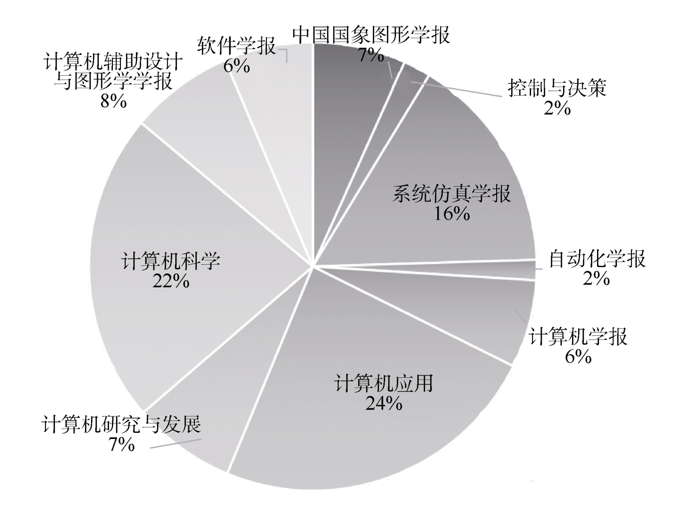


图 1　样本中各个期刊的计算机科学类文献占比

样本中计算机科学类文献数量在 2006 年–2015 年的变化情况如图 2 所示。2006 年–2009 年文献数量相对比较平稳，随后出现一个下降趋势。

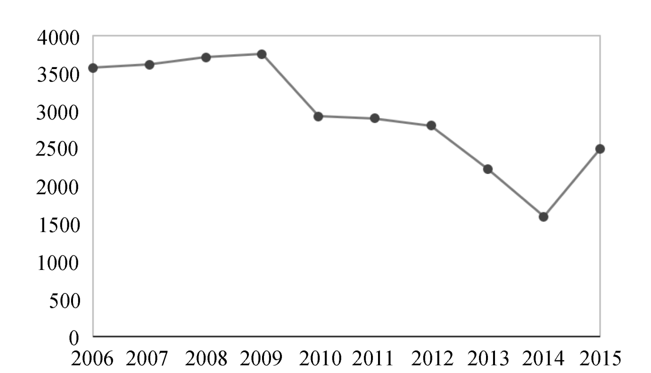


图 2　2006 年–2015 年样本文献年度总量变化趋势

3.2　文献数据预处理与 LDA 参数设置

主题模型的数据输入格式为文档–词矩阵，一行表示一个文档，一列表示一个词。矩阵的条目 m_{ij} 表示第 j 个词汇出现在第 i 个文档中的次数。矩阵的行数等于语料库中的文档数，矩阵的列数等于词汇库中词汇量的大小。因而得到文献数据后，需要对文献进行预处理，以得到文档–词矩阵。本文借用开源软件 R 中的 tm 包对文献数据进行预处理，首先将每篇文献的英文标题、英文关键字和英文摘要分别合并，得到 29 621 个文档；再将文档文本化，形成一个语料，并依次去除标点符号及数字，以及与主题内容无关的停顿词(如 and，then，paper)；最后将语料中的词进行词根化，并删除在少于 3 个文档中出现的词，从而得到一个 29 621 行 10 405 列的文档–词矩阵。

得到文档–词矩阵后，借助开源软件 R 中的 topicmodels 包[23]，构建 LDA 模型。在构建模型前需要确定模型的最优主题数，故设定文档–主题分布 θ 的参数 α=0.1，主题–词分布 φ 的参数 β=0.1，迭代次数 iter=1000，选择 Gibbs Sampling 估计模型的后验参数。首先将主题数 K 依次定为 5-50，发现 K 在 17 与 20 之间，模型的困惑度较低。因此将主题数依次设定为 17、18、19、20，运行 LDA 模型，观察主题之间的语义排他性与主题内部的语义一致性[11]，发现主题数为 18 的主题模型，能较好地涵盖计算机领域的研究内容。因此将主题数 K=18 作为本实验的最优主题数。

LDA 模型的运行结果主要有两个：29 621 篇文献的主题分布 $\theta_{i,j}$，其中 $\theta_{i,j}$ 表示文献 i 中主题 j 的概率；18 个主题的词项分布 $\varphi_{j,v}$，其中 $\varphi_{j,v}$ 表示主题 j 中词 v 的概率。

3.3　实验结果与分析

(1) 文献–主题分布与主题–词分布

根据 LDA 模型的实验结果，得到 29 621 篇文献的主题分布 $\theta_{i,j}$ 和 18 个主题的词项分布 $\varphi_{j,v}$，如表 1 和表 2 所示。$\theta_{1,14}$=0.8417 表明文献 1 的主要研究内容为主题 14，即“检测算法”。根据表 1 可进行主题强度分析，也可对文档进行分类；根据表 2 的高概率主题词，可为每个主题命名，同时分析主题的内容。

表 1　计算机科学类文献的主题分布

$\theta_{i,j}$	1	2	3	4	…	14	15	16	17	18
1	0.0098	0.0009	0.0009	0.0009	…	0.8417	0.0009	0.0009	0.0098	0.0009
2	0.0010	0.0010	0.0317	0.2464	…	0.5634	0.0010	0.0112	0.0112	0.0215
3	0.0010	0.0010	0.0110	0.8527	…	0.0010	0.0210	0.0010	0.0010	0.0010
…	…	…	…	…	…	…	…	…	…	…
29619	0.0007	0.0571	0.0007	0.0078	…	0.4168	0.0007	0.0007	0.0007	0.0007
29620	0.0017	0.0017	0.0017	0.0017	…	0.3027	0.0017	0.0184	0.0017	0.0017
29621	0.0011	0.0011	0.5981	0.0011	…	0.0544	0.0011	0.0011	0.0757	0.0544

根据表 2 的主题–词分布，发现主题内部的词高度相关。如主题“安全密钥”中高概率词 secur、scheme、key、protocol、signatur 和 authent，均与主题“密钥安全”紧密相关。这表明 LDA 模型在提取计算机科学领域文献的潜在主题方面是有效的。

(2) 主题强度分析

主题强度主要描述了主题在某一时期的热门程度，本文中用 θ 表示。在某一时期关于某个主题的文献数量越多，说明该主题的强度越高，可以被认为是热点主题。为了解计算机领域的主题变化模式，笔者将时间“年份”作为变量，将文档–主题分布 θ 按年计算，得出各个主题的主题强度分布情况。根据主题每年的强度大小，可做一个自回归模型，找出主题强度上升的研究主题及主题强度下降的研究主题。本实验

表 2　计算机科学类文献主题的 6 个高概率词

主题标识	主题的 6 个高概率词					
1 安全密钥	secur	scheme	key	protocol	signatur	authent
2 计算系统	system	comput	agent	model	applic	technolog
3 无线传感网络	network	node	rout	sensor	wireless	algorithm
4 图像分割	imag	segment	algorithm	color	region	edg
5 控制系统	model	control	system	simul	predict	time
6 点线面	point	surfac	curv	algorithm	model	mesh
7 图像处理	algorithm	code	watermark	imag	transform	compress
8 并行系统	parallel	perform	data	system	memori	comput
9 特征识别	featur	recognit	classif	vector	face	algorithm
10 数据挖掘	data	algorithm	cluster	queri	tree	set
11 资源调度	schedul	time	resourc	algorithm	task	system
12 系统仿真	simul	system	model	virtual	design	process
13 软件服务	servic	model	web	softwar	system	architectur
14 检测算法	algorithm	object	detect	track	motion	match
15 优化算法	algorithm	optim	search	particl	genet	solv
16 测试模型	model	test	system	net	program	softwar
17 语义信息模型	semant	inform	retriev	model	text	web
18 互联网安全	network	detect	model	trust	evalu	system

(注: 表中的词已经过词根化处理。)

中, 在95%的置信水平下, 18个主题中有7个趋势上升的主题和 6 个趋势下降的主题, 其余 5 个研究主题的趋势变化不明显。表 3 是主题强度发生显著变化的 13 个主题。

表 3　主题强度发生显著变化的主题

主题强度上升的主题		主题强度下降的主题	
主题标签	上升趋势	主题标签	下降趋势
图像分割	0.318*** (0.060)	无线传感网络	–0.140** (0.058)
并行系统	0.218** (0.049)	控制系统	–0.189** (0.065)
特征识别	0.297*** (0.035)	点线面	–0.142** (0.048)
检测算法	0.186** (0.039)	图像处理	–0.167*** (0.027)
优化算法	0.138** (0.035)	资源调度	–0.450** (0.120)
语义信息模型	0.160** (0.037)	软件服务	–0.479*** (0.050)
互联网安全	0.185*** (0.023)		

(注: ***表示 99%的置信水平, **表示 95%的置信水平。)

主题“图像分割”与“特征识别”的主题强度上升较快, 主要源于大数据时代下, 人们越来越注重借助计算机技术对文本进行分析, 以减少相关研究者的工作量。主题“资源调度”与“软件服务”的主题强度下降较快, 但需要注意主题强度下降并不表明该主题不受研究者关注, 只是其受关注程度有下降的趋势。

为观察不同时期的主题强度差异, 可将 2006 年–2015 年划分为两个时间窗口期: 2006 年–2010 年与 2011 年–2015 年。表 4 是不同时期下的 5 个热点主题及其主题强度。据表 4 可知, 2006 年–2015 年期间, 热点主题依次为“图像分割”、“无线传感网络”、“数据挖掘”、“系统仿真”、“检测算法”。同时, 对比不同时间窗口的主题强度, 发现主题“图像分割”、“无线传感网络”、“数据挖掘”的主题强度始终较高; 主题“特征识别”在 2011 年–2015 年期间的主题强度较高, 因而其相对其他主题来说, 可被定义为新兴主题。

表 4　不同时期的 5 个高强度主题

2006-2015		2006-2010		2011-2015	
主题	强度	主题	强度	主题	强度
图像分割	0.0751	系统仿真	0.0741	图像分割	0.0861
无线传感网格	0.0701	无线传感网络	0.0727	数据挖掘	0.0682
数据挖掘	0.0667	图像分割	0.0676	检测算法	0.0663
系统仿真	0.0623	软件服务	0.0665	无线传感网络	0.0663
检测算法	0.0617	数据挖掘	0.0656	特征识别	0.0646

(3) 文献聚类

为更好地了解不同主题的研究状态，根据表 1 中的概率主题分布 $\theta_{i,j}$，对每个主题 j 下的文献数量进行分析。本文设定 θ 的阈值为 0.4[11]，即如果文献 i 在主题 j 中的概率值大于等于 0.4，则文献 i 属于主题 j。表 5 是每个主题下的文献数量及其占比。

表 5　各个主题下的文献数量及占比(θ≥0.4)

主题	文档数	比例
安全密钥	1 144	4.24%
计算系统	925	3.43%
无线传感网络	2 146	7.96%
图像分割	2 213	8.21%
控制系统	1 159	4.30%
点线面	1 417	5.26%
图像处理	1 258	4.67%
并行系统	1 373	5.09%
特征识别	1 449	5.37%
数据挖掘	1 644	6.10%
资源调度	928	3.44%
系统仿真	1 781	6.61%
软件服务	1 595	5.92%
检测算法	1 749	6.49%
优化算法	1 190	4.41%
测试模型	1 508	5.59%
语义信息模型	1 170	4.34%
互联网安全	1 298	4.81%
交叉主题	1 014	3.76%
合计	26 961	100%

根据文献在每个主题的占比，发现主题“图像分割”、“无线传感网络”、“系统仿真”、“检测算法”、“数据挖掘”的文献数量均相对较高，与表 4 的结果趋同，再次表明 LDA 模型适合于挖掘计算机科学类文献的研究主题。

为更好地理解主题的研究内容，可以根据文档的概率主题分布 $\theta_{i,j}$，对各个主题选择与其高度相关的文献。表 6 展示了与热点主题“图像分割”、“无线传感网络”、“系统仿真”、“检测算法”及“数据挖掘”高度相关的 3 个典型文档。通过阅读与主题高度相关的的典型文档，能够更好地把握主题的研究内容。

表 6　对热点主题进行文献举例

主题	代表性文档(作者，年份，论文题目)
图像分割	张建伟等(2013)，《局部熵驱动下的脑 MR 图像分割与偏移场恢复耦合模型》
	刘瑞娟等(2012)，《融合局部和全局图像信息的活动轮廓模型》
	任鸽等(2011)，《基于局部区域拟合模型的磁共振图像分割与偏移估计算法》
无线传感器网络	徐昕等(2010)，《基于链路质量的无线传感器网络任播路由协议》
	郝晓辰等(2009)，《基于路径损耗的无线传感器网络分布式拓扑控制算法》
	李小亚等(2008)，《一种异构传感器网络的能量有效路由算法》
系统仿真	丁海燕等(2009)，《基于 HLA 的舰空导弹反导仿真系统的设计与实现》
	赵旭东等(2009)，《ARJ21 飞机工程模拟器关键技术研究》
	张禹等(2008)，《水下滑翔机器人实时仿真平台研究与开发》
检测算法	李伟生等(2014)，《基于时空背景模型的自适应运动目标检测方法》
	孟苑等(2008)，《基于运动点积累的视频运动目标提取》
	王哲等(2008)，《一种基于立体视觉的运动目标检测算法》
数据挖掘	郭鑫等(2011)，《动态数据库中的频繁子树挖掘算法》
	田卫东等(2008)，《基于简化分辨矩阵的粗糙集属性约简算法》
	陈明等(2006)，《一种有效的基于图的关联规则挖掘算法》

(注：本文定义若 $\theta_{i,j}$ > 0.95，则高度相关。)

4　结　语

本文借助 LDA 主题模型，结合模型困惑度和对主题内容的经验判断来确定模型的最优主题数，同时考虑文献的发表时间，针对 29 621 篇计算机科学文献挖掘出 18 个潜在主题，并对这 18 个主题的内容和强度进行研究，通过分析找到 7 个趋势上升的主题和 6 个趋势下降的主题。根据 LDA 模型输出的 29 621 个文档的概率主题分布，设定主题概率阈值，将文献分配到各个主题下进行数量统计，并对热点主题进行文档列举，细致了解热点主题的研究内容。结果表明，LDA 模型能够较为准确地提取计算机科学文献的研究主题，有利于科研人员对该学科领域的发展状态进行初步了解，同时把握未来的研究方向，寻找新兴主题。

当然，本文亦存在不足之处：

(1) 本文选取的样本数量较大(29 621 篇文献)，具

有一定的实践意义，但仅考虑国内计算机科学领域的专业期刊，未考虑国内学者在国内综合类优秀期刊及国外优秀期刊上发表的计算机科学类文献，未来研究可考虑扩大样本容量进行主题分析，以充分了解我国计算机科学领域的发展状态;

(2) LDA 模型假定主题之间相互独立，而同一个学科领域的主题之间往往存在不可分割的联系，因此未来研究中，可以将相关主题模型(Correlated Topic Model)[24]的思想加入到模型中;

(3) 本文仅考虑发表时间这一个外部特征，未来研究中可考虑借助结构主题模型(Structural Topic Model)[25-27]，加入作者特征、期刊类别等外部信息，更精确地了解一个学科领域的研究状态。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Blei D M. Probabilistic Topic Models [J]. Communications of the ACM, 2012, 55(4): 77-84.

[3] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences, 2004, 101(S1): 5228-5235.

[4] 郭玉，蔚海燕. 我国计算机科学发展态势文献计量分析[J]. 计算机应用研究，2007, 24(12): 28-31. (Guo Yu, Yu Haiyan. Biblio-metrilogical Analysis on Development Trends of Computer Science in China [J]. Application Research of Computers, 2007, 24(12): 18-31.)

[5] 陈国良，孙广中，徐云，等. 并行计算的一体化研究现状与发展趋势[J]. 科学通报，2009，54(8): 1043-1049. (Chen Guoliang, Sun Guangzhong, Xu Yun, et al. Integrated Research of Parallel Computing: Status and Future [J]. Chinese Science Bulletin, 2009, 54(8): 1043-1049.)

[6] 章锦文，马远良. 神经网络计算机的现状与发展趋势[J]. 计算机科学，1993，20(6): 24-27. (Zhang Jinwen, Ma Yuanliang. The Development Situation and Direction of Neurocomputer [J]. Computer Science, 1993, 20(6): 24-27.)

[7] Zheng B, McLean D C, Lu X. Identifying Biological Concepts from a Protein-related Corpus with a Probabilistic Topic Model[J]. BMC Bioinformatics, 2006, 7(4): 58.

[8] Hall D, Jurafsky D, Manning C D. Studying the History of Ideas Using Topic Models [C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008: 363-371.

[9] Wu H, Wang M, Feng J, et al. Research Topic Evolution in "Bioinformatics"[C]. In: Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE). IEEE, 2010: 1-4.

[10] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation [J]. Journal of the American Society for Information Science and Technology, 2011, 62(1): 185-204.

[11] Piepenbrink A, Nurmammadov E. Topics in the Literature of Transition Economies and Emerging Markets [J]. Scientometrics, 2015, 102(3): 2107-2130.

[12] 贺亮，李芳. 科技文献话题演化研究[J]. 现代图书情报技术，2012(4): 61-67. (He Liang, Li Fang. Topic Evolution in Scientific Literature [J]. New Technology of Library and Information Service, 2012(4): 61-67.)

[13] 关鹏，王曰芬，傅柱.不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析[J]. 图书情报工作，2016，60(2): 112-121. (Guan Peng, Wang Yuefen, Fu Zhu. Effect Analysis of Scientific Literature Extraction Based on LDA Topic Model with Different Corpus [J]. Library and Information Service, 2016, 60(2): 112-121.)

[14] 李湘东，张娇，袁满.基于 LDA 模型的科技期刊主题演化研究[J]. 情报杂志，2014，33(7): 115-121. (Li Xiangdong, Zhang Jiao, Yuan Man. On Topic Evolution of Scientific Journal Based on LDA Model [J]. Journal of Intelligence, 2014, 33(7): 115-121.)

[15] 王曰芬，傅柱，陈必坤.采用 LDA 主题模型的国内知识流研究结构探讨：以学科分类主题抽取为视角[J]. 现代图书情报技术，2016(4): 8-19. (Wang Yuefen, Fu Zhu, Chen Bikun. Analyzing Knowledge Structure Research with LDA Model[J]. New Technology of Library and Information Service, 2016(4): 8-19.)

[16] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报，2011，30(6): 583-590. (Wang Ping. Literature Knowledge Mining Based on Probabilistic Topic Model [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(6): 583-590.)

[17] 叶春蕾，冷伏海. 基于引文—主题概率模型的科技文献主题识别方法研究[J]. 情报理论与实践，2013，36(9): 100-103. (Ye Chunlei, Leng Fuhai. Discovering the Topic of Science Literature Based on Citation-Topic Model [J]. Information Studies: Theory & Application, 2013, 36(9): 100-103.)

[18] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书情报工作，2014，58(22): 70-77. (Wang Ping.Topic Extraction and Evolution for Scientific Literature Based on

Hierarchical Probabilistic Topic Model [J]. Library and Information Service, 2014, 58(22): 70-77.)

[19] 王金龙, 徐从富, 耿雪玉. 基于概率图模型的科研文献主题演化研究[J]. 情报学报, 2009, 28(3): 347-355. (Wang Jinlong, Xu Congfu, Geng Xueyu. Study on Research Topic Evolution Based on Probabilistic Graphical Models [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(3): 347-355.)

[20] 李湘东, 廖香鹏, 黄莉. LDA 模型下书目信息分类系统的研究与实现[J]. 现代图书情报技术, 2014 (5): 18-25. (Li Xiangdong, Liao Xiangpeng, Huang Li. Research and Implementation of Bibliographic Information Classification System in LDA Model [J]. New Technology of Library and Information Service, 2014 (5): 18-25.)

[21] 秦晓慧, 乐小虬. 基于 LDA 主题关联过滤的领域主题演化研究[J]. 现代图书情报技术, 2015 (3): 18-25. (Qin Xiaohui, Le Xiaoqiu. Topic Evolution Research on a Certain Field Based on LDA Topic Association Filter [J]. New Technology of Library and Information Service, 2015 (3): 18-25.)

[22] 杨如意, 刘东苏, 李慧. 一种融合外部特征的改进主题模型[J]. 现代图书情报技术, 2016(1): 48-54. (Yang Ruyi, Liu Dongsu, Li Hui. An Improved Topic Model Integrating Extra-Features [J]. New Technology of Library and Information Service, 2016 (1): 48-54.)

[23] Grün B, Hornik K. Topicmodels: An R Package for Fitting Topic Models [J]. Journal of Statistical Software, 2011, 40(13): 1-30.

[24] Blei D M, Lafferty J D. A Correlated Topic Model of Science [J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[25] Roberts M E, Stewart B M, Tingley D, et al. The Structural Topic Model and Applied Social Science [J]. Medical Journal of Australia, 2013, 155(6): 419-420.

[26] Roberts M E, Stewart B M, Tingley D. stm: R Package for Structural Topic Models [J]. General Information, 2014, 57(1): 445-460.

[27] Roberts M E, Stewart B M, Tingley D, et al. Structural Topic Models for Open-Ended Survey Responses [J]. American Journal of Political Science, 2014, 58(4): 1064-1082.

作者贡献声明:

杨海霞: 提出研究思路, 设计研究方案, 论文撰写及最终版本修订;
高宝俊: 提出研究思路, 设计研究方案, 提出论文修改建议, 论文最终版本修订;
孙含林: 采集、清洗和分析数据, 提出论文修改建议。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: haixiayang@whu.edu.cn。

[1] 杨海霞, 高宝俊, 孙含林. 2006-2015 计算机科学文献数据.csv. 文献数据.

收稿日期: 2016-06-02
收修改稿日期: 2016-07-31

Extracting Topics of Computer Science Literature with LDA Model

Yang Haixia Gao Baojun Sun Hanlin
(Economics and Management School, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper employs text mining technology to automatically identify research topics from large amounts of scientific literature and then detects future trends. [**Methods**] First, we used the LDA model to find both topical prevalence and contents of articles published by the top ten computer science journals in China. Second, we described the evolution of major topics with the help of publishing dates. [**Results**] We extracted 18 topics from 29, 621 computer science papers and then identified 7 trending topics as well as 6 less popular ones. [**Limitations**] Our study did not include papers published overseas by Chinese authors. [**Conclusions**] The proposed method could help us learn the evolution of computer science research and then grasp the emerging trends.

Keywords: Computer science LDA Topic mining Topic prevalence Document cluster